KTH Computer Science
and Communication

Exam in DD2431 Machine Learning
2014-10-31, kl 14.00 – 18.00

Aids allowed: *calculator*, *language dictionary*.

A Questions for pass or fail

Only one alternative is correct for each question.

Note: To pass the exam you must give the correct answer on almost *all* questions in this section. Only *one* error will be accepted, so be very careful not to make any unnecessary mistakes here.

A-1 Naive Bayes Classifier

What is the underlying assumption unique to a *naive Bayes* classifier?

a) Prior probabilities are available

b) The number of features (the dimension of feature space) is large

c) A Gaussian distribution is assumed for the feature values

d) All features are regarded as independent

Solution: d

A-2 Probabilistic Learning

Given two events A and B, what does it mean if the conditional probability $P(A|B)$ is equal to the probability of A, that is $P(A|B) = P(A)$?

a) event A is more probable than event B

b) event B is more probable than event A

c) events A and B are independent

d) observing B gives information about A

Solution: c

A-3 Overfitting

You have trained a model (classifier) using some training sample data. Under which conditions is overfitting most likely to occur?

a) Relatively complex model is used where few training samples are available

b) Relatively complex model is used where many training samples are available

c) Relatively simple model is used where few training samples are available

d) Relatively simple model is used where many training samples are available

Solution: a

A-4 Learning Theory

In learning theory, what is the meaning of the term *hypotheses space*?

a) The initial guess used to initialize an incremental learning algorithm

b) The distance between positive and negative training samples

c) The set of all possible solutions to the learning task

d) Any assumption about what algorithm to use for learning

Solution: c

A-5 Support Vector Machines

What is a *Support Vector Machine*?

a) A method to compute similarity between vectors

b) A classifier which uses borderline samples to define the separation surface

c) A machine based learning algorithm which aggregates the result from multiple learners

d) An algorithm for compressing vector data using weighted averages

Solution: b

A-6 The Subspace Method

Which statement best describes the characteristics of the subspace method for classification?

a) To repsesent each class by a low-dimensional subspace

b) To exploit that input patterns are uniformity distributed across the input space

c) To generate a high-dimensional subspace that covers all the classes

d) To ensure an equal number of positive and negative samples

Solution: a

A-7 Classification and Regression

Which one of those statements represents the output formats of classification and regression?

a) They are both real-valued

b) They are both discrete

c) Real-valued for classification and discrete for regression

d) Real-valued for regression and discrete for classification

Solution: d

A-8 Decision Trees

What principle is commonly used when building a decision tree?

a) Choose features that maximize information gain

b) Use optimal weights for individual training samples

c) Maximize the tree height

d) Minimize the number of leaf nodes

Solution: a

B Questions for higher grades

Preliminary number of points required for different grades:

$$\begin{aligned} 22 \le p \le 24 &\rightarrow A \\ 18 \le p < 22 &\rightarrow B \\ 12 \le p < 18 &\rightarrow C \\ 6 \le p < 12 &\rightarrow D \\ 0 \le p < 6 &\rightarrow E \end{aligned}$$

B-1 Terminology (4p)

For each term (a–h) in the left list, find the explanation from the right list which best describes how the term is used in machine learning.

a) Support Vector

b) Negative sample

c) k-means

d) Normal distribution

e) Posterior probability

f) The Lasso

g) Principal Component Analysis

h) k-nearest neighbour

1) Data without useful information

2) Probability divided by the bias

3) Class prediction by a majority vote

4) Probability distribution of the most common samples

5) Probability after observation

6) Samples on the margin of the decision surface

7) Evidence for a specific hypothesis

8) Sudden increase of information

9) Training data which is not part of the concept being learned

10) Clustering method based on centroids

11) Continuous PDF defined by mean vector and covariance matrix

12) An approach to regression that results in variable selection

13) Probability at a later time

14) An unsupervised method for dimensionality reduction

15) Method for estimating the mean of k observations

16) The last solution

Solution: a-6, b-9, c-10, d-11, e-5, f-12, g-14, h-3

B-2 Probability based learning (4p)

You work in a pharmaceutical company that is developing a new test to diagnose a certain disease. You know that the disease is relatively rare and affects about 0.1% of the population. You have the task of defining the requirements for the test in terms of probability of false alarms (positive test on healthy patients) and false rejections (negative test on sick patients). In the requirements these two probabilities are assumed to be equal to each other and called α. In order for the test to be diagnostically useful two conditions must be verified:

1. if the test is positive, the patient should be more likely to be sick than healthy.

2. if the test is negative, the patient should be more likely to be healthy than sick.

What range of values for α verifies these conditions?

Solution: The first constraint on α comes from the rules on probabilities: $0 \leq \alpha \leq 1$

We call S and $\bar{S}$ the events of the patient being sick and healthy, respectively. Similarly we call T the event of a positive response from the test and $\bar{T}$ the negative response. From the problem we have:

Priors:

$$\begin{aligned} P(S) &= 0.001 \\ P(\bar{S}) &= 1 - P(S) = 0.999 \end{aligned}$$

Likelihoods:

$$\begin{aligned} P(T|\bar{S}) &= P(\bar{T}|S) = \alpha \\ P(\bar{T}|\bar{S}) &= P(T|S) = 1 - \alpha. \end{aligned}$$

The two conditions defined by the problem correspond to posterior probabilities:

1. after observing a positive response from the test, we want the posterior probability of the patient being sick to be greater than that of being healthy, in formulas: $P(S|T) > P(\bar{S}|T)$, or, equivalently, $P(S|T) > 0.5$,

2. similarly, given a negative response, we want $P(\bar{S}|\bar{T}) > P(S|\bar{T})$.

Focusing on condition 1, we can expand both sides of the inequality using Bayes formula:

$$P(S|T) = \frac{P(T|S)P(S)}{P(T)} > \frac{P(T|\bar{S})P(\bar{S})}{P(T)} = P(\bar{S}|T)$$

where $P(T)$ is equal for both terms and can be disregarded in the comparison, as long as it is non-zero[1]. Substituting $P(T|\bar{S}) = \alpha$ and $P(T|S) = 1 - \alpha$, we obtain:

$$(1 - \alpha)P(S) > \alpha P(\bar{S})$$

that, remembering that $P(S) + P(\bar{S}) = 1$, is verified for:

$$\alpha < P(S) = 0.001.$$

With similar derivations for condition 2 we obtain $\alpha < P(\bar{S}) = 0.999$. The range of values that satisfies all the above constraints on α is:

$$0 \leq \alpha < 0.001$$

[1] An extra careful solution, would have proved that this is verified for the current problem parameters

B-3 Classification (3p)

Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20% on the training data. We also get the average error rate (averaged over both test and training data sets) of 25%. Next we use 1-nearest neighbor and get an average error rate (averaged over both test and training data sets) of 18%.

a) What was the error rate with 1-nearest neighbor on the training set?

b) What was the error rate with 1-nearest neighbor on the test set?

c) Based on these results, indicate the method which we should prefer to use for classification of new observations, with a simple reasoning.

Solution: a) 0% b) 36% c) Logistic regression because it achieves lower error rate on the test data, 30%.

B-4 Information Content (3p)

Consider a game where you throw two dice, one red and one green. If the red die has a smaller number, you lose; if the green die has a smaller number, you win. If they are equal, it is a draw.

a) How unpredictable is the outcome of this game (win, lose or a draw)? Answer in terms of entropy, measured in bits.

b) After first throwing the red die and seeing the result, what is the unpredictability of the outcome of the game in terms of entropy?

c) What is the expected information gain from seeing the result of first throw?

Note: if you do not have a calculator, do answer with an expression, but simplify as much as possible.

Solution:
a) Let $f(p_1, p_2, p_3) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - p_3 \log_2 p_3$
Entropy: $f(15/36, 6/36, 15/36) \approx 1.483$

b) Six possible outcomes with same probability.
$e_1 = f(\frac{0}{6}, \frac{1}{6}, \frac{5}{6})$ $e_2 = f(\frac{1}{6}, \frac{1}{6}, \frac{4}{6})$ $e_3 = f(\frac{2}{6}, \frac{1}{6}, \frac{3}{6})$ $e_4 = f(\frac{3}{6}, \frac{1}{6}, \frac{2}{6})$ $e_5 = f(\frac{4}{6}, \frac{1}{6}, \frac{1}{6})$ $e_6 = f(\frac{5}{6}, \frac{1}{6}, \frac{0}{6})$
The unpredictability:$(e_1 + e_2 + e_3 + e_4 + e_5 + e_6)/6 \approx 1.120$

c) The information gain will be approximately: $1.483 - 1.120 = 0.363$

B-5 Regression with regularization (2p)

In ridge regression, relative to least squares, a term called *shrinkage penalty* is added in the quantity to be minimised.

a) What is the effect of having this term on regression coefficients?

b) Ridge regression will give improved prediction accuracy in some situations. Briefly explain when this happens in terms of bias-variance trade-off.

Solution: a) The regression coefficients are supressed, and shrink towards zero as the impact of the penalty is increased. b) When the increase in bias is less than the decrease in variance.

B-6 Ensemble of Decision Trees (2p)

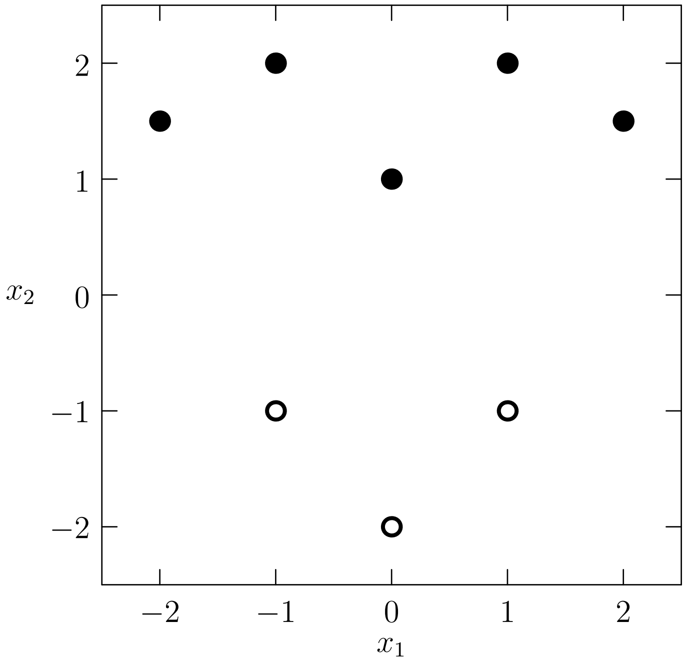
Suppose we produce ten bootstrapped samples (bootstrap replicates) from a data set containing red and green classes. We then apply a classification tree to each bootstrapped sample and, for a specific value of x, produce 10 estimates of $P(Red|x)$: 0.4, 0.75, 0.3, 0.55, 0.4, 0.7, 0.65, 0.65, 0.75, and 0.4.

There are two common ways to combine these results together into a single class prediction. One is (i) the majority vote approach, and the other is (ii) based on the average probability. In this example, explain what the final classification is under each of these two approaches.

Solution: Average probability: The average of the probabilities is more than 50%, thus Red.
Majority vote: The number of red predictions is greater than that of green predictions, thus Red.

B-7 Support Vector Machine (3p)

Given the training data illustrated in the figure. Filled circles are positive examples, unfilled are negative.



a) What will the support vectors be for a corresponding (linear) support vector machine?

b) Estimate the α values for each support vector when a linear kernel

$$\mathcal{K}(\vec{x}, \vec{y}) = \vec{x}^T \cdot \vec{y} + 1$$

is used.

Hint: Do not attempt to solve the optimization problem! Make use of the fact that you already know the support vectors and that you know what the value of the indicator function must be at these points.

Solution:

Support vectors (points touching the widest possible separating band):

$$x_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad x_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \qquad x_3 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Indicator function is:

$$\text{ind}(\vec{x}) = \sum_i \alpha_i t_i \mathcal{K}(\vec{x}, \vec{x}_i) = \alpha_1 \mathcal{K}(\vec{x}, \vec{x}_1) - \alpha_2 \mathcal{K}(\vec{x}, \vec{x}_2) - \alpha_3 \mathcal{K}(\vec{x}, \vec{x}_3)$$

Use the fact that the indicator function is -1 and 1 at the two margins. Therefore:

$$\text{ind}(\vec{x}_1) = 1 \qquad \text{ind}(\vec{x}_2) = -1 \qquad \text{ind}(\vec{x}_3) = -1.$$

Entering the actual support vector coordinates gives:

$$\alpha_1 \cdot 2 + \alpha_2 \cdot 0 + \alpha_3 \cdot 0 = 1$$

$$\alpha_1 \cdot 0 + \alpha_2 \cdot 3 + \alpha_3 \cdot 1 = -1$$

$$\alpha_1 \cdot 0 + \alpha_2 \cdot 1 + \alpha_3 \cdot 3 = -1$$

This gives:

$$\alpha_1 = 0.5 \qquad \alpha_2 = 0.25 \qquad \alpha_3 = 0.25$$

B-8 Learning Theory (3p)

The term *hypotheses* is used when analyzing concept learning in learning theory. For each one of the learning algorithms below, describe what constitutes a hypothesis.

a) Decision tree learning based on information gain

b) Support vector machines

c) Logistic regression

Solution:

a) One specific decision tree

b) A set of support vectors and their corresponding α-values

c) A set of weights

Good Luck!